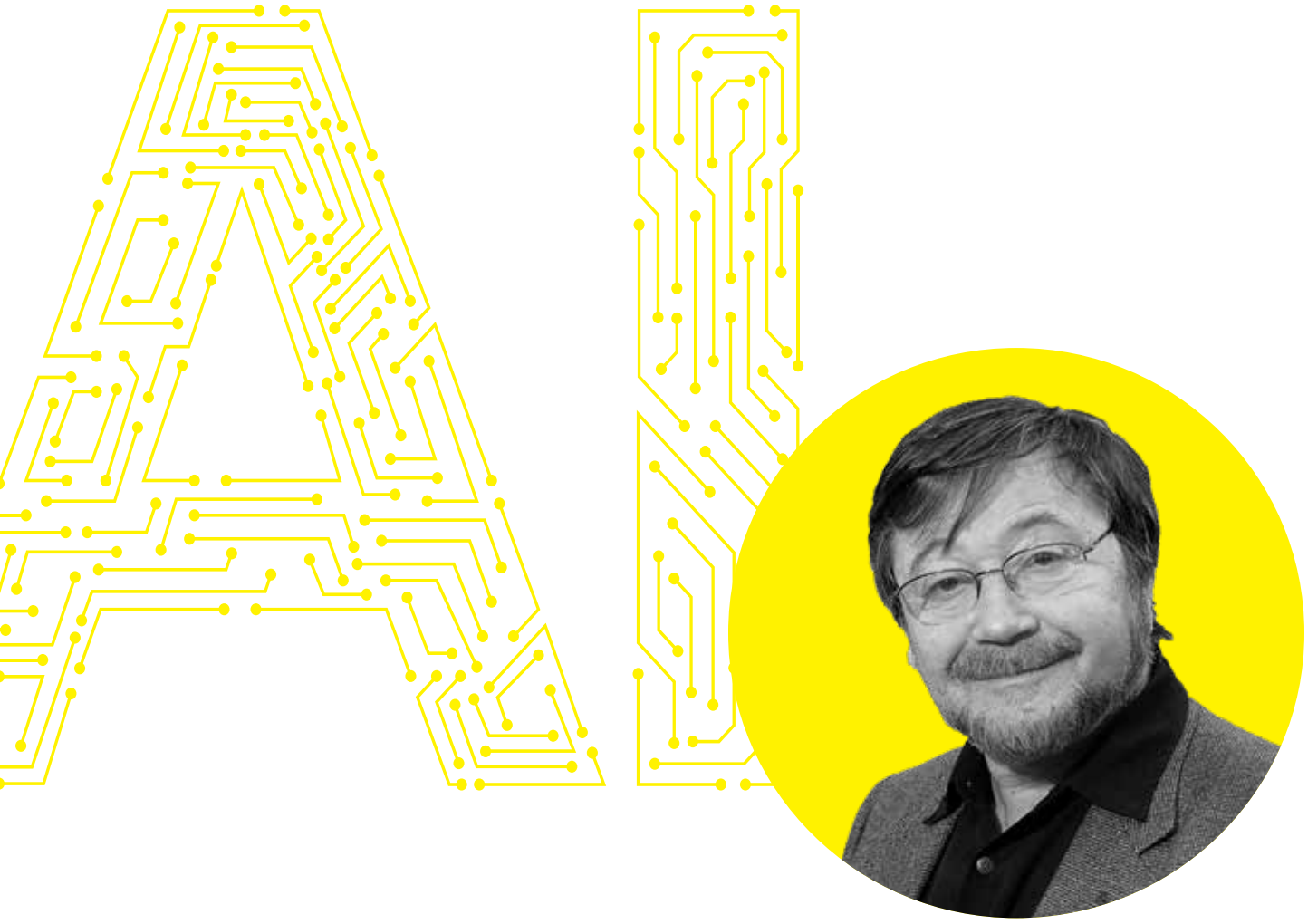

Pearl

JUDEA PEARL, AI, and CAUSALITY: WHAT ROLE DO STATISTICIANS PLAY?

In the first half of 2023, the machine learning programs ChatGPT and GPT-4 changed the landscape of artificial intelligence research seemingly overnight. Judea Pearl's research bridges the subjects of statistics and artificial intelligence and highlights the importance of causality in both settings. Dana Mackenzie, Pearl's co-author for *The Book of Why,* interviews him here to get his take on recent developments. When they wrote their book in 2018, Pearl contended machine learning had not yet moved past the first rung of the "ladder of causation." Computers could not correctly answer queries about interventions and still less about counterfactual scenarios. Has his assessment changed?

MACKENZIE: Can you tell me your first reactions to ChatGPT and GPT-4? Did you find their capabilities surprising?

PEARL: Aside from being impressed, I have had to reconsider my proof that one cannot get any answer to any causal or counterfactual query from observational studies. What I didn't take into account is the possibility that the text in the training database would itself contain causal information. The programs can simply cite information from the text without experiencing any of the underlying data.

For example, I asked it the questions about the firing squad [from Chapter 1 of *The Book of Why*], such as what would have happened to the (now deceased) prisoner if rifleman 1 had refrained from shooting. At first it goes into side tracks and tells you, for example, "it is dangerous to shoot people." But if you have time and prompt it correctly, it will get closer to the correct answer: "If soldier 1 refrained from shooting after receiving the signal, the prisoner could still have been killed by soldier 2, assuming he received and acted upon the same signal." Finally, it gives an A+ answer: "Given the additional information, if each soldier always fires upon receiving a signal and any one soldier's shot is enough to cause the prisoner's death, then the prisoner would still be dead if soldier 1 refrained from shooting. This is because soldier 2, following the captain's signal, would have fired his shot, causing the prisoner's death. This is an example of 'overdetermination' in causation, where an effect (the prisoner's death) has more than one sufficient cause (either soldier's shot)."

Here, I have to make a cautionary note. In spite of its impressive command of vocabulary, ChatGPT doesn't have a causal model itself. It doesn't have a structure into which it can imbed new knowledge. If you ask it about another problem with the same causal structure, say about inoculations, you'll have to prompt it again from scratch. It won't generalize.

MACKENZIE: Is it doing better than previous AIs have?

PEARL: Which ones do you mean? If they tried to do deep learning from data of actual firing squads, not from texts about causal relationships, then they could not even understand the question, let alone give a coherent answer.

MACKENZIE: Is this a new world of AI, even for you?

PEARL: Yes, it's a new one. It's similar to a world with causal information you can learn from teachers who cannot experiment for themselves but learned from teachers who learned from books. You can learn a lot of causal information from books. We [humans] are still different, because we have an innate causal model or an innate template into which we are born and which we periodically update with new information.

Causal reasoning is not all you need for human-like AI. You have other components, like natural language processing and vision, that are also necessary for artificial general intelligence (AGI). It is in this one little corner of causal inference that we have been successful at achieving deep understanding by combining models and data, an understanding that can be generalized to other areas of AI.

MACKENZIE: I'd like to turn now to something *Amstat News* readers will be curious about: What can statisticians contribute to AI research?

PEARL: I once said every statistician is a frustrated philosopher, struggling to extract

Dana Mackenzie is a mathematician who became a science journalist. He has written *The Book of Why* with co-author Judea Pearl, as well as popular science and math articles for publications such as *Science, New Scientist, Smithsonian, Discover*, and *American Scientist*. He lives in Santa Cruz with his wife, dog, cat, and random foster kittens.

Never in history has there been such an acceleration of the speed of evolution.

meaning from data. Statisticians are brought up to believe all knowledge comes from data and, since they are experts on data processing, they must also be experts in the philosophy of knowledge (epistemology).

But as I just said, to understand the world of causes and effects, you need to combine models and data, a rather neglected exercise in mainstream statistics. Once we open statistics to modern vocabulary, including causal and counterfactual relationships, we open the door for statisticians to participate in current issues faced by AI researchers as well as philosophers of science.

Even those who wish to adhere to standard statistical vocabulary can contribute appreciably to causal inference tasks. In causal inference, we distinguish between estimands and estimates; the former being distributional expressions of what needs to be estimated, and the latter being the actual estimates obtained from finite samples of a distribution. This distinction defines a symbiotic division of labor between statisticians and causal inference researchers, respectively. Some of the estimands produced by causal analysis may seem strange to statisticians. An example is the estimand produced by the front-door criterion [See *Book of Why,* Chap. 7.]. Addressing them through the lens of modern estimation techniques should be a challenging endeavor for creative statisticians. This is something they do well, and we need their ingenuity. But if they want to know where the estimand came from, causal modeling would be necessary.

MACKENZIE: In *The Book of Why,* we said current AI programs operate at the first level of the ladder of causation, the level of observation or "fitting functions to data." Has this changed?

PEARL: It has. The ladder restrictions [e.g., level-two queries cannot be answered by level-one data] do not hold anymore because the data is text, and text may contain information on levels two and three.

MACKENZIE: In particular, does reinforcement learning make it possible for a machine to understand level two on the ladder of causation by giving it data on interventions?

PEARL: Yes, that is correct. I would say it's at level one and three-fourths. Reinforcement learning trains machines on interventions. For example, you can train them on chess. They can decide, after playing many games, that a certain move will give them a higher probability of checkmate than another move. However, they cannot infer from this anything about a third move they haven't tried. They also cannot combine interventions to infer what will happen if they do both A and B. For that, again, you would need a causal model.

MACKENZIE: That leads to my next question. How can you tell whether you have the right causal model?

PEARL: That is the central question of epistemology in general. We never know for sure. We can only falsify models but cannot prove they are correct.

MACKENZIE: I remember this exact question came up when we were on the podcast for *Science* magazine. The interviewer asked us how you know whether you have the right model and you gave the most wonderful two-word answer: "By argument." Can you explain what you meant?

PEARL: I don't remember that question! But "by argument" is how you form a consensus in the society of scientists. That's how theories become accepted. The development of science has two parts. First is testing your theory. We know now when a causal model can be tested, and we know what observations or experiments to conduct in order to (potentially) falsify it. The second component is to try out a modification. If you have a causal model, modify it and try out another model, a refinement of the old one.

That's what science is all about. Einstein doesn't completely throw out Newtonian physics—it's still in there, but he refines it by making a local perturbation.

Can a machine perform a local perturbation? Not today. But I can envision how it can be done. A machine that decides what experiments to perform next should also be able to modify its theory and continue to progress. That's how I think general AI will eventually become smarter than scientists.

MACKENZIE: Even AI researchers agree we need ethical guidelines for the use of AI. What guidelines would you recommend?

PEARL: I have to answer this question at two different levels. First, at the level of ChatGPT, it's already dangerous because it can be misused by dictators or by greedy businesses to do a lot of harm: combining and distorting data, using it to control a segment of the population. That can be done even today with ChatGPT. Some regulation is needed to make sure the technology doesn't fall to people who will misuse it, even though it's in the very early stage of development. It's not general AI yet, but it still can be harmful.

The second danger is when we really have general AI, machines that are a million times more powerful [than humans]. At this point I raise my hands and say we don't even have the metaphors with which to understand how dangerous it is and what we need to control it.

I used to feel safe about AI. What's the big deal? We take our chances with teenagers, who think much faster than us. Once in a while we make a mistake and we get a Putin, and the world suffers. But most of the time, education works. But with AI, we are talking about something totally different. Your teenagers are now a hundred million times faster than you, and they have access to a hundred million times larger space of knowledge. Never in history has there been such an acceleration of the speed of evolution. For that reason, we should worry about it, and I don't know how to even begin to speak about how to control it.

MACKENZIE: But didn't we talk about this in *The Book of Why?* We discussed the concept of regret, the idea that a machine with a causal model could compare what happened with what would have happened if it took a different course of action. Do you still think regret can equip a machine to make its own ethical judgements?

PEARL: Regret and responsibility will of course be part of AGI and will be implemented eventually using counterfactual logic. Where it will go, I don't know. No matter how well we program the guards of responsibility for this new species, it might decide it wants to dominate the world on its own. It happened to *Homo sapiens*. We extinguished all the other forms of human, the Neanderthal and *Homo erectus*. Imagine what a machine 10 million times smarter could do. It's unbelievable.

The idea of dominating the world could be one of those local perturbations I talked about. The machine might try it out, decide it's fun, and pursue it with vigor.

MACKENZIE: So are you pessimistic now about giving AIs human-compatible ethics fast enough?

PEARL: You can try to form a committee to regulate it, but I don't know what that committee will do.

MACKENZIE: To conclude the interview, do you have any predictions about what we are going to see in AI in the next year or five years?

PEARL: Do you want to ask me what we are going to see, or what I'd like to see? I'd like to see a shift in emphasis from machine learning to general AI. ChatGPT actually slowed down our progress toward general AI. More and more of our resources will be poured into that direction and not into the correct way of doing AI.

MACKENZIE: But maybe that's a good thing. You said general AI is something to worry about.

PEARL: Here, I am torn. Maybe it's a blessing that ChatGPT is so stupid and society is so intoxicated with it. So maybe we are safe from the danger of creating the new species I mentioned. ■